CS 134: Networks	Lecture 2, Wednesday, January 25
Prof. Yaron Singer	Spring 2017

1 Introduction

In this course we will be interested in understanding the structure of networks we observe.

Scientific question: *Why do the networks we observe in the real world look the way they do?*

In Figure 1 we illustrate a network of romantic relationships between high school students, and in Figure 2 we depict friendship relationships amongst high school students where the nodes are color-coded according the students' race. What are the kind of processes that generate them? Can we predict the structure of networks, at any level? For example, how many components there will be? What features of networks should we think are surprising/context-specific, and which are implied by very general "physical laws" that map across settings?

Modeling networks. One of the biggest challenges in network analysis is developing tractable models. In general, networks are *dynamic* objects driven by both the *underlying structure* of the situation and some element of *randomness*. For example, social networks are driven by both strategic considerations of the individuals involved (let's apply to this school to hang out with smart kids) and by some element of chance (I just met this cool kid in the dog park).

We will start by separating these two drivers of networks; we will first study purely random models of network formation and then more structured models, including ones where agents make choices about how to link. The random models can identify processes that generate certain features but do not explain those processes might arise. In a strategic model, the explanation for a specific feature of a network is explicitly tied to choices.

We will begin by thinking about the special case of *static random networks*. These are networks in which all nodes are established at the same time and in which links are drawn between them according to some probabilistic rule.

2 Networks as graphs

- We use *graphs* to model networks;
- A graph is an ordered pair $G = (V, E)$ of a set of vertices (nodes) V and edges $E \subseteq V \times V$. Graphs can be directed, undirected, edge-weighted, node-weighted.

Graphs are a powerful modeling tool that allow us to analyze real-world networks. As discussed above, an important modeling tool that we wish to introduce is that of *random graphs*. But before we introduce the concept of random graphs, it is worth discussing what are the advantages of using randomness.

Probabilistic modeling. When should we use probabilistic models?

- Maybe the world is genuinely unpredictable (no amount of data would suffice to nail deterministic predictions);
- Maybe the modeler does not have access to the data he would need; random modeling captures all the unobserved variables;
- Probabilistic models as a way of "handling our ignorance".

Deterministic vs. random approach.

- **Deterministic Approach:**
  - collect data (age, sex, survey responses);
  - predict what the network will look like (output = one predicted network);
  - give yourself an A if it looks exactly as you predicted, give yourself an F otherwise.
- **Random Approach**
  - collect data;
  - for each possible network, predict the *probability* that it forms;
  - look at the *distribution* of some network statistics and see if you match them *on average*.

Rule of thumb: *simpler models (fewer parameters) are generally better.*

3 Random Graphs

Before we introduce and define random graphs, it will be useful to describe an analogy. The analogy that we will use is that of a die.

3.1 A die is a random variable

We can denote $\mathcal{D}(n)$ as the set of all possible outcomes for the role of a die with n faces. For example, for a die with 6 faces, the possible outcomes are $\{1, 2, \ldots, 6\}$. For those of you who played $D\&D$ you well know that a die does not necessarily need to have 6 faces. It can have 3 or 20 faces, or any integral number (in theory). A die with n faces is simply a *random variable* that takes values in $\mathcal{D}(n)$. Alternatively, we can say that a die is a distribution on $\mathcal{D}(n)$. One example of a die is a *fair die*, where every face has an equal probability of occurring. But if you ever visited Vegas, you know that a die need not be fair.

Figure 1: Romantic relationship among high school students (Add Health dataset).

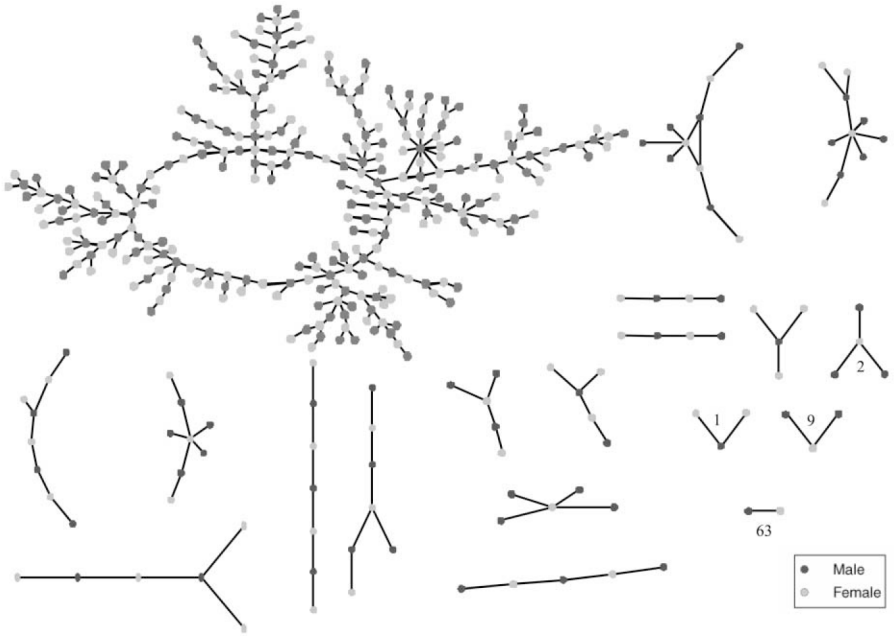
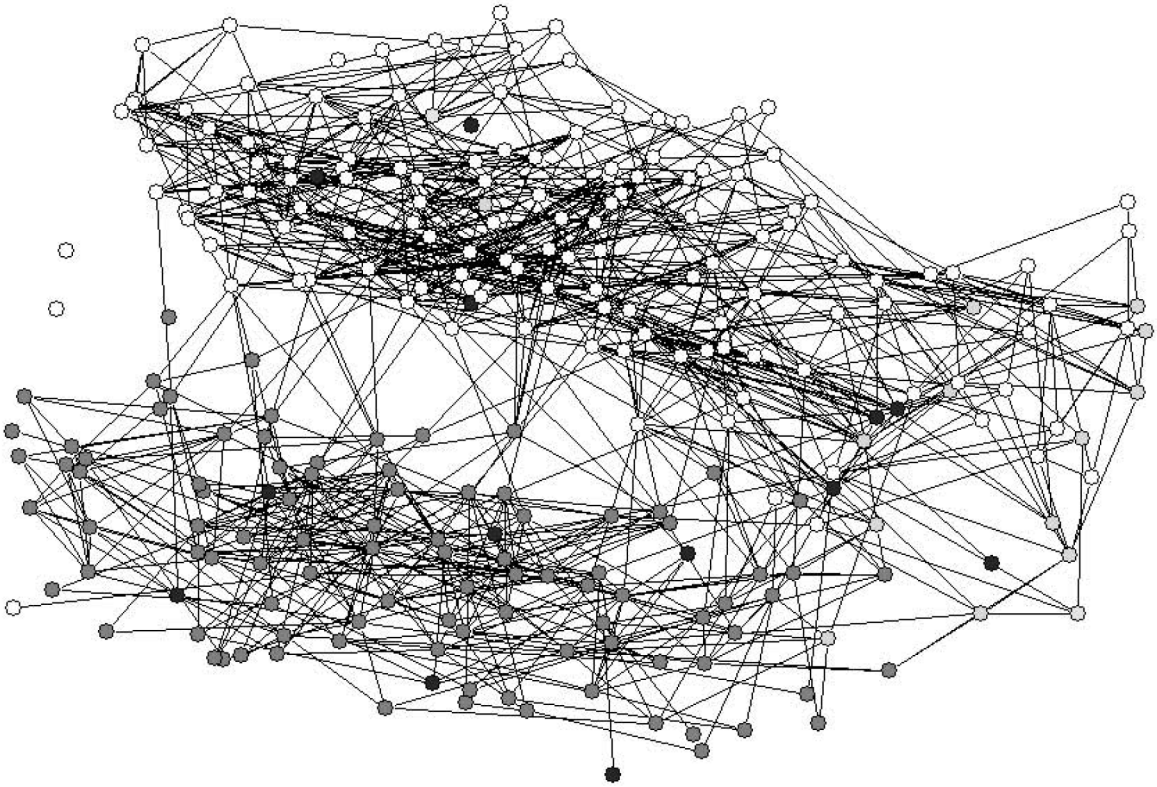


Figure 2: Friendships among high school students coded by race (Add Health dataset).



3.2 A random graph is a random variable

We can denote $\mathcal{G}(n)$ as the set of all possible graphs with n nodes. A random graph is then a random variable that takes values in $\mathcal{G}(n)$, or alternatively, a random graph is a *distribution* on $\mathcal{G}(n)$. One example of a random graph is the Erdös-Rényi $G(n,p)$ model.

3.3 Erdös-Rényi $G(n,p)$ model

Let $\mathcal{G}(n)$ be the set of all (undirected) graphs on n nodes. The *random graph* $G(n,p)$ is a random variable that takes values in $\mathcal{G}(n)$. To define $G(n,p)$, fix a set of nodes $V = \{1,\ldots,n\}$, and let a link between any two nodes i and j be included in the set E of edges with probability $p \in [0,1]$, where the formation of links is independent. This is a binomial model of link formation[1], which gives rise to a manageable set of calculations regarding the resulting network structure.

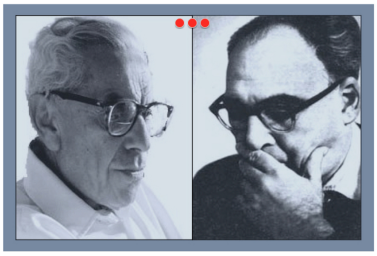


Figure 3: Paul Erdös & Alfréd Rényi.

To return to our analogy, the outcome of the roll of a standard die is a random variable taking values in $\{1,\ldots,6\}$. So $G(n,p)$ is not a fixed graph but the outcome of a random experiment. The random variable has two parameters:

- *Size*, or number of nodes n
- *Density*, or probability p that a link is formed between any two nodes.

3.4 Example: a realization of $G(4,0.5)$

- Draw nodes a, b, c and d.
- There are $\binom{4}{2} = \frac{4\times 3}{2} = 6$ possible links.
- For each potential link, flip a coin, with probability $p = 0.5$ of landing heads. Include the link if the coin comes out heads.

3.5 Analysis of the Erdös-Rényi Model

Example 1. *For instance, if $n = 3$, then*

- *a complete network forms with probability p^3;*
- *any given network with two links (there are 3 such networks) forms with probability $p^2(1-p)$;*
- *any given network with one link forms with probability $p(1-p)^2$;*
- *the empty network forms with probability $(1-p)^3$.*

[1] On the binomial distribution, see Section 4.5.4 of Jackson's book

More generally, if we fix a particular graph $\widetilde{G} \in \mathcal{G}(n)$ on the node set $V = \{1, 2, \ldots, n\}$, what is $\mathbb{P}[G(n,p) = \widetilde{G})$, the probability that a draw of $G(n,p)$ turns out to equal $\widetilde{G}$ exactly? For the equality event to happen, the draw of $G(n,p)$ has to have all edges present in $\widetilde{G}$, and none of the edges absent in $\widetilde{G}$. Observe that in a network on n nodes there are $m = \frac{n(n-1)}{2}$ potential links. So the probability of observing a network $\widetilde{G}$ that has i specific links of the m possible links is[2]

$$\mathbb{P}[G(n,p) = \widetilde{G}] = p^i(1-p)^{m-i}$$

of forming under this process: the p^i corresponds to the i links present in $\widetilde{G}$ being present in the draw of $G(n,p)$, and the $(1-p)^{m-i}$ corresponds to the $m-i$ links absent in $\widetilde{G}$ being absent in the draw of $G(n,p)$.

3.5.1 Degree Distribution

We can calculate some statistics that describe the network. For instance, we can find the degree distribution fairly easily. The degree of a node is the number of links that a node has. In a random graph on the node set $V = \{1, 2, \ldots, n\}$, this is a random variable. Let d_i denote the (random) degree of a node i. This random variable can take any value in $\{0, 1, 2, \ldots, n-1\}$, (note i can't have a self-link, hence the $n-1$). The probability that $i \in V$ has exactly d links is

$$\mathbb{P}[d_i = d] = \binom{n-1}{d} p^d (1-p)^{n-1-d} \tag{1}$$

To see this:

- Node i can potentially form a link with $n-1$ nodes,
- Consider any set S of d nodes that does not contain node i, i.e. $S \subseteq V \setminus \{i\}$. The probability that node i has a link with each of the nodes in this set and with no other nodes is $p^d(1-p)^{n-1-d}$
- There are $\binom{n-1}{d}$ different sets $S \subseteq V \setminus \{i\}$ of d nodes that do not contain node i.

4 Bonus section: Approximation via Poisson Distribution

For large n and small p, this binomial expression is approximated by a Poisson distribution, so that the probability that a node has d links is approximately[3]

$$\frac{e^{-(n-1)p}((n-1)p)^d}{d!}.$$

[2]Note that there are $\binom{m}{i}$ distinct graphs with a total of i edges on on n nodes, but each has a distinct set of i edges. Thus, the probability of a graph with a specific i edges is as shown in the notes, and the probability that a graph has i total edges is $\mathbb{P}[G(n,p] \text{ has } i \text{ edges}) = \binom{m}{i} p^i (1-p)^{m-i}$.

[3]To see this, note that for large n and small p, $(1-p)^{n-1-d}$ is roughly $(1-p)^{n-1}$. Write

$$(1-p)^{n-1} = \left(1 - \frac{(n-1)p}{n-1}\right)^{n-1}$$

which, if $(n-1)p$ is either constant or shrinking (if we allow p to vary with n), is approximately $e^{-(n-1)p}$. Then, for fixed d, large n and small p, $\binom{n-1}{d}$ is roughly $\frac{(n-1)^d}{d!}$.

If we let $\mu = (n-1)p$ denote the expected degree of a node,[4] we see that the formula becomes

$$\mathbb{P}[d_i = d] \approx e^{-\mu}\frac{\mu^d}{d!}.$$

Computing the probability that $d_i = d$ corresponds to the following exercise: we fix our gaze on a given i and draw $G(n,p)$ "many times," and we are curious about how often $d_i = d$ across those draws. There is another thing we are curious about: in a *single* draw of $G(n,p)$—a single random graph—*what fraction of nodes* will have degree d? Note this is a different question. Rather than looking across many draws, we are fixing a draw and looking across many nodes. However, it turns out that the answer is essentially the same: the *fraction* of nodes with degree d is approximately

$$\frac{e^{-(n-1)p}((n-1)p)^d}{d!}.$$

Why? Because the d_i, as i ranges over V, are (nearly) independent random variables. If we have a random variable d_i with a given probability distribution function, $f(d) = \mathbb{P}[d_i = d]$, and we take many independent copies of it $(d_1, d_2, \ldots, d_n)$, then the empirical distribution of the many draws is very likely to be very close to f.[5]

Given the approximation of the degree distribution by a Poisson distribution, the class of random graphs for which each link is formed independently with equal probability is often referred to as the class of *Poisson random networks*.

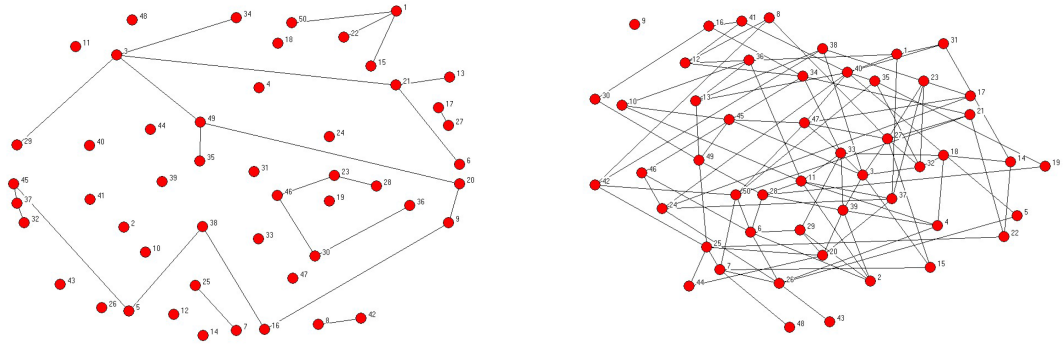
4.1 Illustrations

To provide a better feel for the structure of such networks, consider a couple of Poisson random networks for different values of p. Set $n = 50$ nodes, as this number produces a network that is easy to visualize.

- **Figure 4(a)**: Let us start with an expected degree of 1 for each node, which is equivalent to setting p at roughly .02. Based on the approximation of a Poisson distribution, we should expect 37.5 percent of the nodes to be isolated, which is between 18 and 19 nodes. There happen to be 19 isolated nodes in this network.
- **Figure 4(b)**: Let us increase the probability to $p = .078$ which is roughly the threshold at which isolated nodes disappear. Based on the approximation of a Poisson distribution we should expect about 2 percent of the nodes to be isolated or roughly 1 out node out of 50. There happens to be 1 isolated node in this network.
- **Figure 5**: compares the realized frequency distribution of degrees with the Poisson approximation. The distributions match fairly closely.

[4]The degree of a node i is the sum of the $n-1$ Bernoulli random variables corresponding to whether i is linked to each of the $j \neq i$. There are $n-1$ of these, and each of them has probability p of being on.

[5]Why "nearly" independent? Note that even though links are formed independently, there is some correlation in the degrees of various nodes, which affects the distribution of nodes that have a given degree. For instance, if $n = 2$, then it must be that both nodes have the same degree. But as the number of nodes n becomes large, however, the correlation of degree between any two nodes vanishes, as the possibility of a link between them is only 1 out of the $n-1$ that each might have. (It is a good exercise to show this.) Thus, as n becomes large, the fraction of nodes that have d links will approach (1).



(a) (b)

Figure 4: Randomly generated network with different probabilities of each link: $p = 0.02$ in (a) and $p = 0.08$ in (b).

- **Figure 6**: The realized frequency distribution of degrees is again similar to the Poisson approximation.

You will have the chance to plot several frequency distributions in your homework.

4.2 Thresholds and Phase Transitions

Degrees are interesting, but often we care about some other aspect of a graph. Will it be connected? Will it have a large component, so that there's the potential for a disease to spread from one person to many others? We would like to understand these properties of a random graph. Next week we will continue our analysis of the $G(n, p)$ model and argue that it is not suitable to explain small-world phenomena that we observe in the real world. Another method of analysis is by making statements about graphs whose number of nodes tend to infinity. For making these kinds of arguments the Poisson approximation is a useful tool and we will see these kinds of arguments later in the course.

4.3 Concluding Thoughts on the Erdös-Rényi Model

- It's the simplest introduction to the important class of probabilistic models of networks.
- But it's missing a lot features we observe in social networks:
  - No dynamics
  - No clustering – my friends are likely to be friends in real friendship networks
  - No homophily – people's connections depend on the *types* of their friends
  - Degree distribution is unrealistic –Poisson is actually unusual in social networks

We will study small-world networks, power-law networks, later in the course.

Figure 5: Frequency distribution of of the randomly generated network in Figure 4 and the Poisson approximation for a probability of .02 on each link. *The units of the $x-$ axis are Degree $+1$ and not Degree*

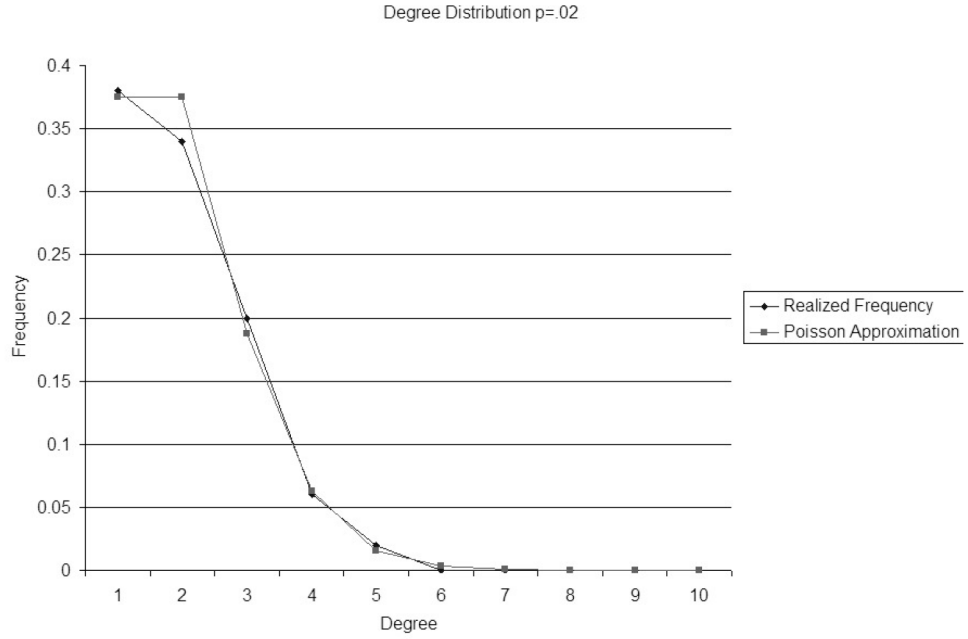


Figure 6: Frequency distribution of the randomly generated network in Figure 4 and the Poisson approximation for a probability of .08 on each link. *The units of the $x-$ axis are Degree $+1$ and not Degree*